Capstone Project-2

Mechanical Fitting Failure Classification

Industry: Gas Pipeline

Prepared by: Prashant Sanghal

Project Motivation:

- Code of Federal Regulations (49 CFR Parts 191, 192) requires gas distribution pipeline operators to report hazardous leaks Involving mechanical fitting (DOT Form PHMSA F-7100.1-2).

- Oldest fitting installation dates back to 1851, 165 years ago.

So, using data let's find out:

What caused the leak and when?
Where and how often?
Was it the same fitting from the same manufacturer?
Who were the Pipeline operators ?

Benefit to Clients:

- Pipeline operators can avoid environmental hazards.
- Manufacturers can re-design better fittings.
- Insurance companies can estimate premiums.
- We can set up pre-leak alert system.
- Reduce customer downtime.
- Transport gas safely to various locations.

Data Story:

- Dataset was available on Kaggle
- File format: Comma separated values, CSV
- Data size: Observations 85,611, Columns 54.
- Numerical values: None, and had to be converted from text.
- Missing values Treated:
  - Initial Phase : 18,789
  - Final Phase : 155,047 (in multiple columns)

Question before us:

How can we preserve critical information?

- 15 columns out of 22, had missing values > 70%.
- Available data in missing columns was critical and had unique values.
- It described leak features important for our study:
  - Was weld defect due to material or design?
  - Was excavation damage at the time of leak discovery or before?
  - Were there natural forces that caused leak or not?

How can we bring this all together?

Missing Dashboard:

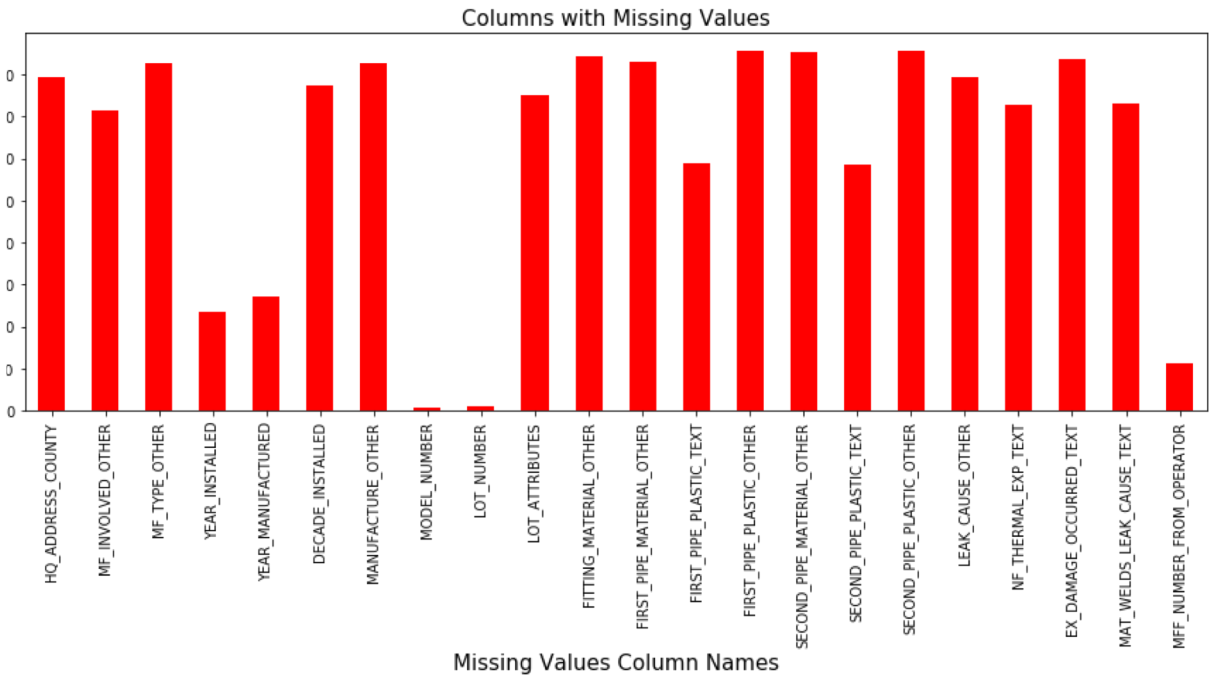
- A function which returns a summary table.
- Showing count of missing values, missing percentage, unique and available values in missing columns.

Dashboard:

```
Average (base) Missing Percentage: 0.73
Number of Missing Columns
(Missing Percentage at 0.7): 15
```

	missing_values	missing_percentage	unique_values_in_missing_columns	available_values_in_missing column
HQ_ADDRESS_COUNTY	79345	0.93	84	6266
MF_INVOLVED_OTHER	71434	0.83	731	14177
MF_TYPE_OTHER	82557	0.96	317	3054
DECADE_INSTALLED	77247	0.90	10	8364
MANUFACTURE_OTHER	82683	0.97	196	2928
LOT_ATTRIBUTES	83343	0.97	723	2268
FITTING_MATERIAL_OTHER	84411	0.99	52	1200
FIRST_PIPE_MATERIAL_OTHER	82845	0.97	41	2766
FIRST_PIPE_PLASTIC_OTHER	85557	1.00	24	54
SECOND_PIPE_MATERIAL_OTHER	85274	1.00	53	337
SECOND_PIPE_PLASTIC_OTHER	85568	1.00	16	43
LEAK_CAUSE_OTHER	79267	0.93	778	6344
NF_THERMAL_EXP_TEXT	72766	0.85	2	12845
EX_DAMAGE_OCCURRED_TEXT	83503	0.98	2	2108
MAT_WELDS_LEAK_CAUSE_TEXT	73060	0.85	2	12551

Distribution of Missing Values in Dataset

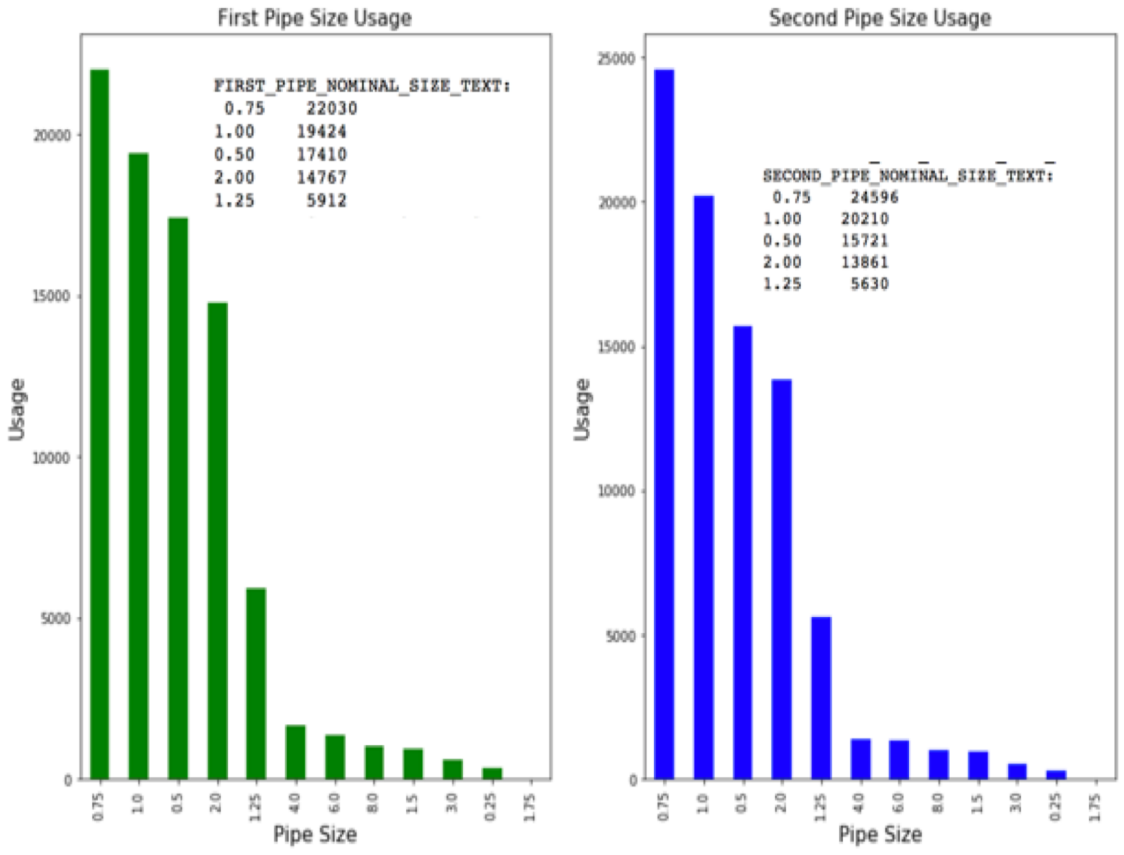


Benefits:

- Quickly scanned 15 missing columns with > 70% missing values.
- Aided in deciding which missing values to keep Vs let go.
- As a result, replaced **18,789** missing values in 9 columns.

Exploratory Analysis:

Most Used Pipe Sizes

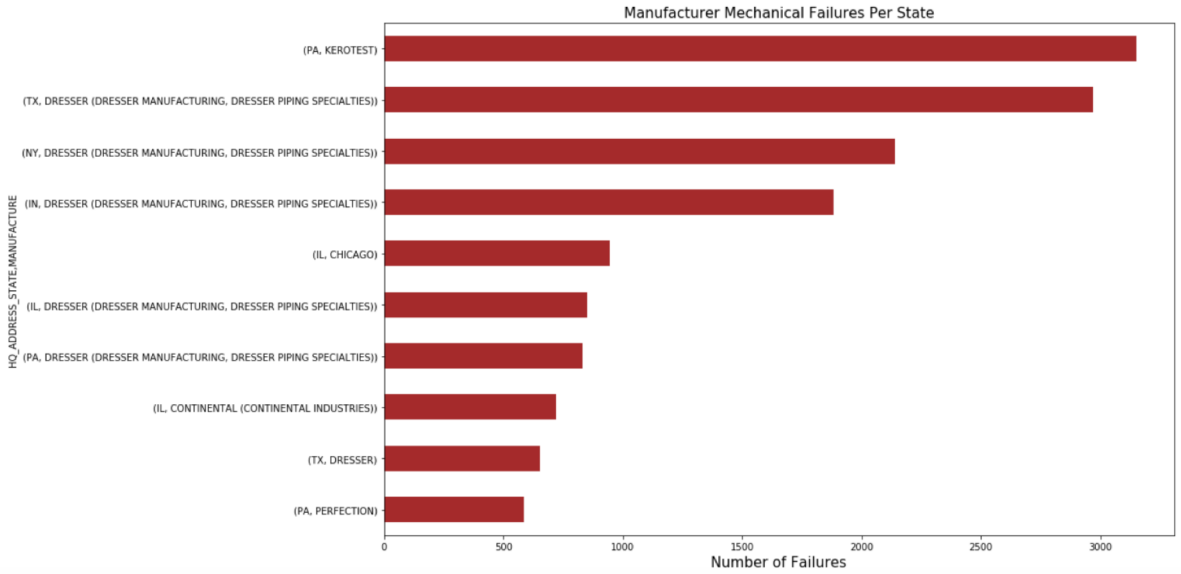


- Pipe size 0.75 showed most usage among pipeline operators.

The Top 10 Known Manufacturers by State where Leak Occurred

HQ_ADDRESS_STATE	MANUFACTURE	
PA	KEROTEST	3152
TX	DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)	2972
NY	DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)	2138
IN	DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)	1881
IL	CHICAGO	944

Name: MANUFACTURE, dtype: int64



- PA, Kerotest had maximum leaks

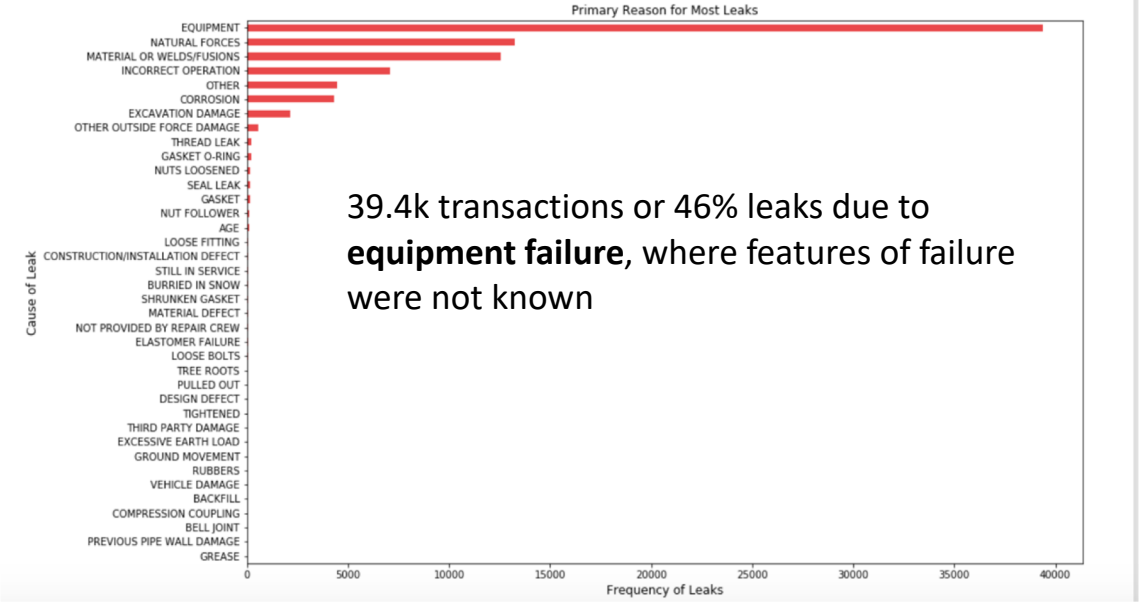
...contd.

Exploratory Analysis:

Top 15 Reasons for the Leak

****Top 15 reasons for leak****

	Reason_Count	% Reason_Count
EQUIPMENT	39370	45.987081
NATURAL FORCES	13230	15.453622
MATERIAL OR WELDS/FUSIONS	12551	14.660499
INCORRECT OPERATION	7070	8.258285
OTHER	4464	5.214283
CORROSION	4330	5.057761
EXCAVATION DAMAGE	2123	2.479822
OTHER OUTSIDE FORCE DAMAGE	575	0.671643
THREAD LEAK	225	0.262817
GASKET O-RING	223	0.260481
NUTS LOOSENED	180	0.210253
SEAL LEAK	163	0.190396
GASKET	144	0.168203
NUT FOLLOWER	133	0.155354
AGE	90	0.105127



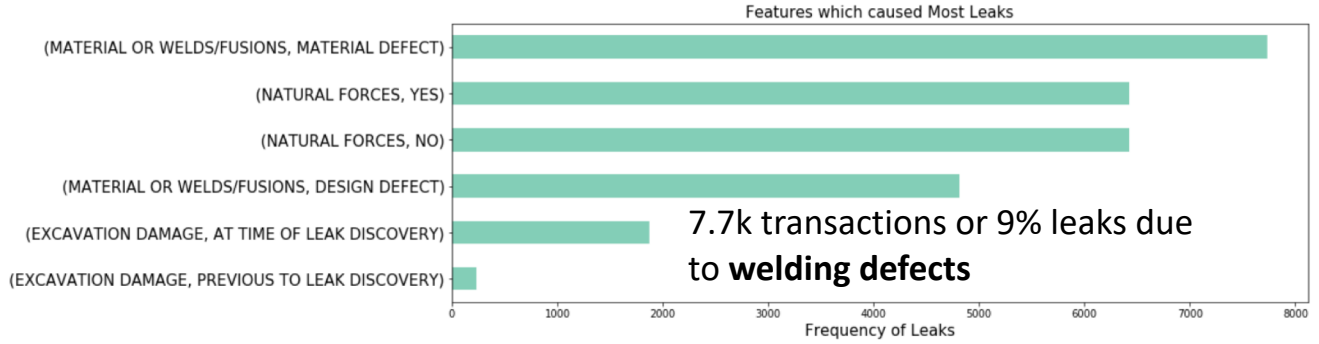
39.4k transactions or 46% leaks due to **equipment failure**, where features of failure were not known

Features which Caused Leak in the Pipeline

****Features which caused leak****

		Feature_Count
LEAK_CAUSE_TEXT	ADDITIONAL_LEAK_FEATURES	
MATERIAL OR WELDS/FUSIONS	MATERIAL DEFECT	7734
NATURAL FORCES	YES	6424
	NO	6421
MATERIAL OR WELDS/FUSIONS	DESIGN DEFECT	4817
EXCAVATION DAMAGE	AT TIME OF LEAK DISCOVERY	1872
	PREVIOUS TO LEAK DISCOVERY	236

		%_Feature_Count
LEAK_CAUSE_TEXT	ADDITIONAL_LEAK_FEATURES	
MATERIAL OR WELDS/FUSIONS	MATERIAL DEFECT	9.033886
NATURAL FORCES	YES	7.503709
	NO	7.500204
MATERIAL OR WELDS/FUSIONS	DESIGN DEFECT	5.626613
EXCAVATION DAMAGE	AT TIME OF LEAK DISCOVERY	2.186635
	PREVIOUS TO LEAK DISCOVERY	0.275666



7.7k transactions or 9% leaks due to **welding defects**

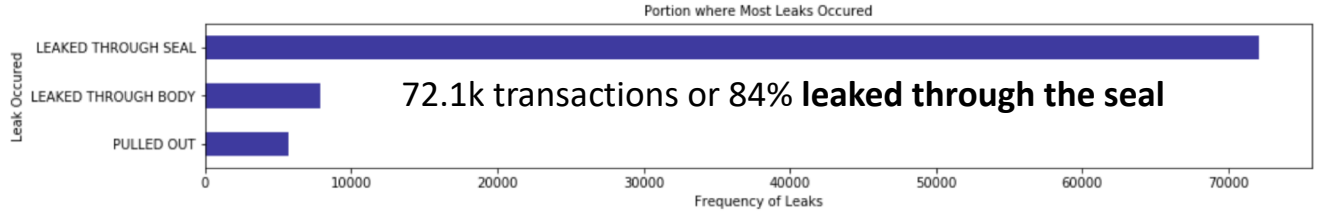
...contd.

Exploratory Analysis:

Portion where most leaks occurred

****Portion where leak occured****

	Occurred_Count	% Occurred_Count
LEAKED THROUGH SEAL	72062	84.173763
LEAKED THROUGH BODY	7867	9.189240
PULLED OUT	5682	6.636998

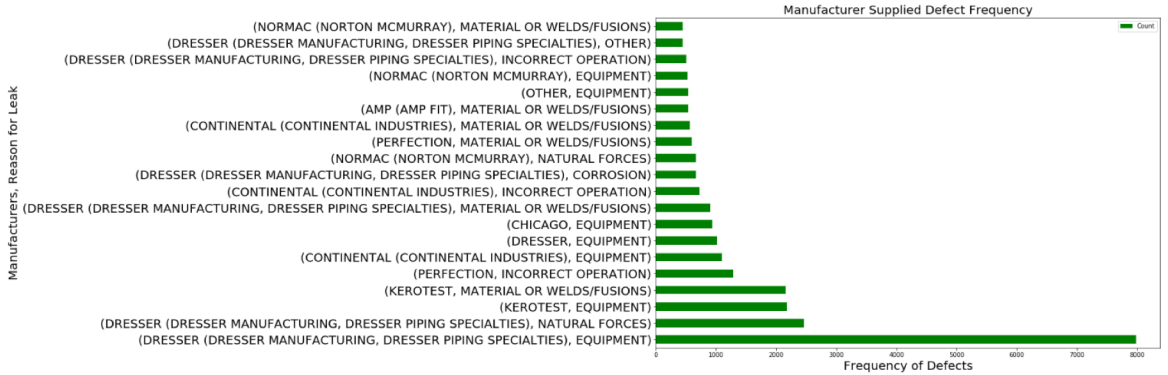


72.1k transactions or 84% **leaked through the seal**

Manufacturer Supplied Defects

Manufacturer Supplied Defect Frequency:

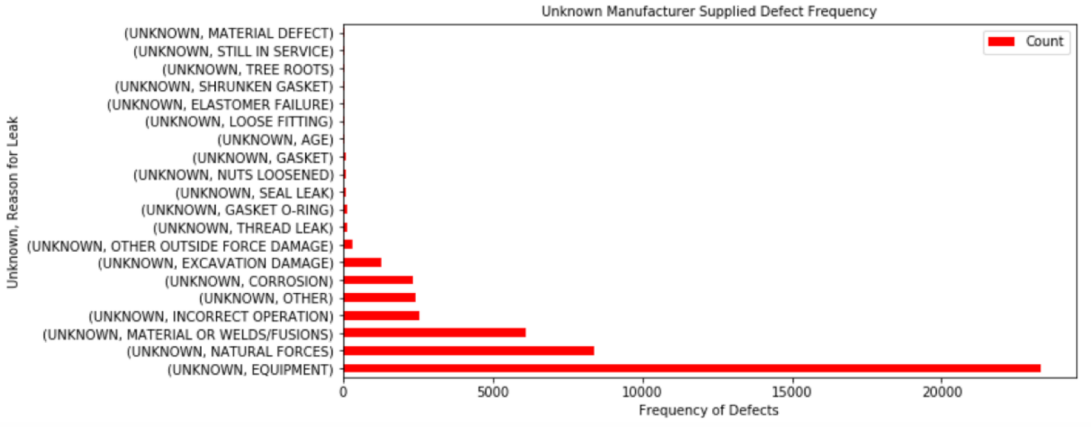
MANUFACTURE	LEAK_CAUSE_TEXT	Count
DRESSER (DRESSER MANUFACTURING, DRESSER PIPING ...	EQUIPMENT	7985
	NATURAL FORCES	2461
KEROTEST	EQUIPMENT	2177
	MATERIAL OR WELDS/FUSIONS	2162
PERFECTION	INCORRECT OPERATION	1285



Unknown Manufacturer Supplied Defects

Unknown Manufacturer Supplied Defect Frequency:

MANUFACTURE	LEAK_CAUSE_TEXT	Count
UNKNOWN	EQUIPMENT	23328
	NATURAL FORCES	8408
	MATERIAL OR WELDS/FUSIONS	6108
	INCORRECT OPERATION	2537
	OTHER	2434



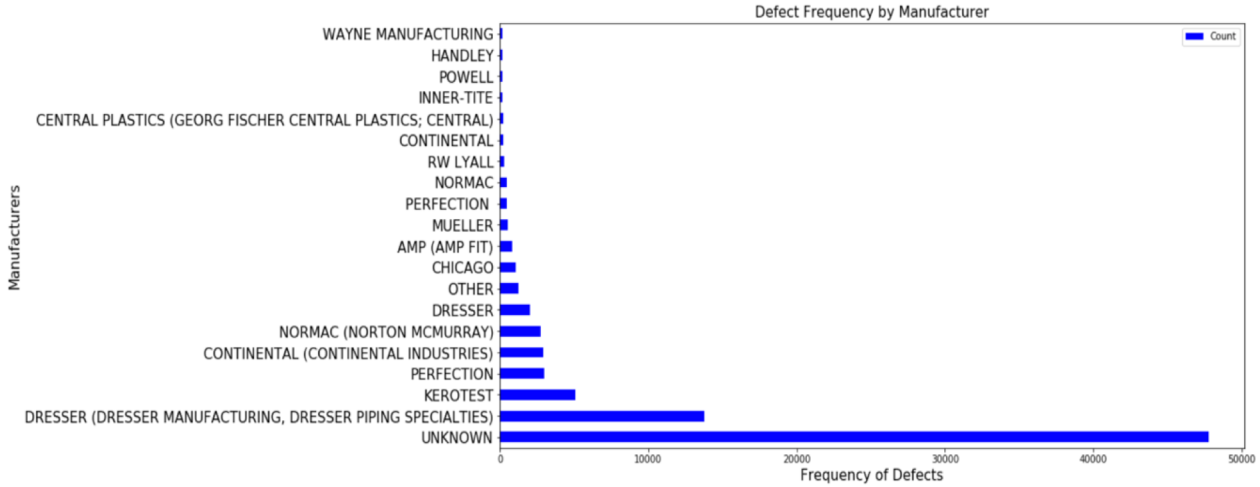
...contd.

Exploratory Analysis:

Defect Frequency by Manufacturers:

Defect Frequency by Manufacturer:

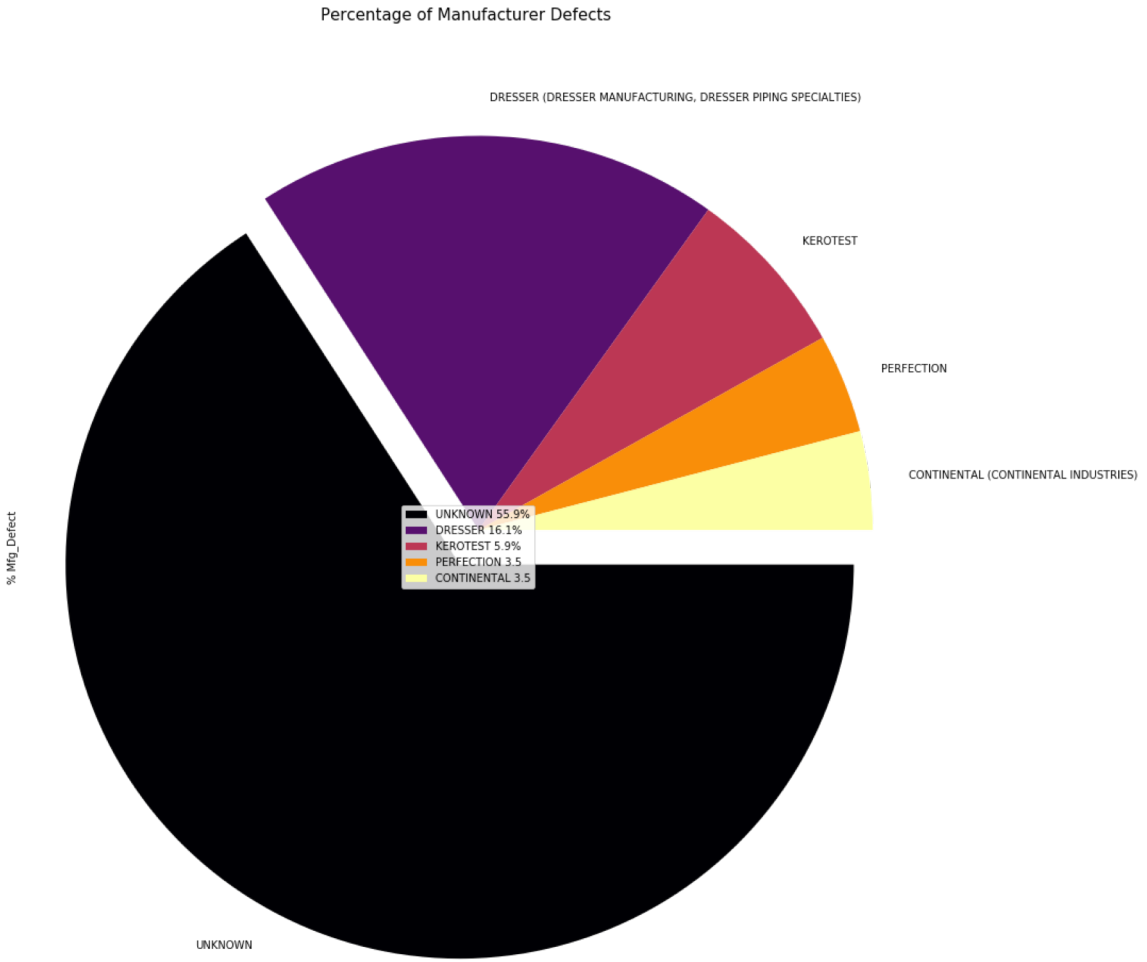
MANUFACTURE	Mfg_Count	% Mfg_Defect
UNKNOWN	47816	55.887234
DRESSER (DRESSER MANUFACTURING, DRESSER PIPING ...	13808	16.138760
KEROTEST	5073	5.929311
PERFECTION	2954	3.452629
CONTINENTAL (CONTINENTAL INDUSTRIES)	2928	3.422240



Observation:

- Unknown: 56% equipment failure.
- Dresser: 16% equipment, natural forces.
- Kerotest: 6% equipment, welding failures.

Percentage of Defects by Manufacturers

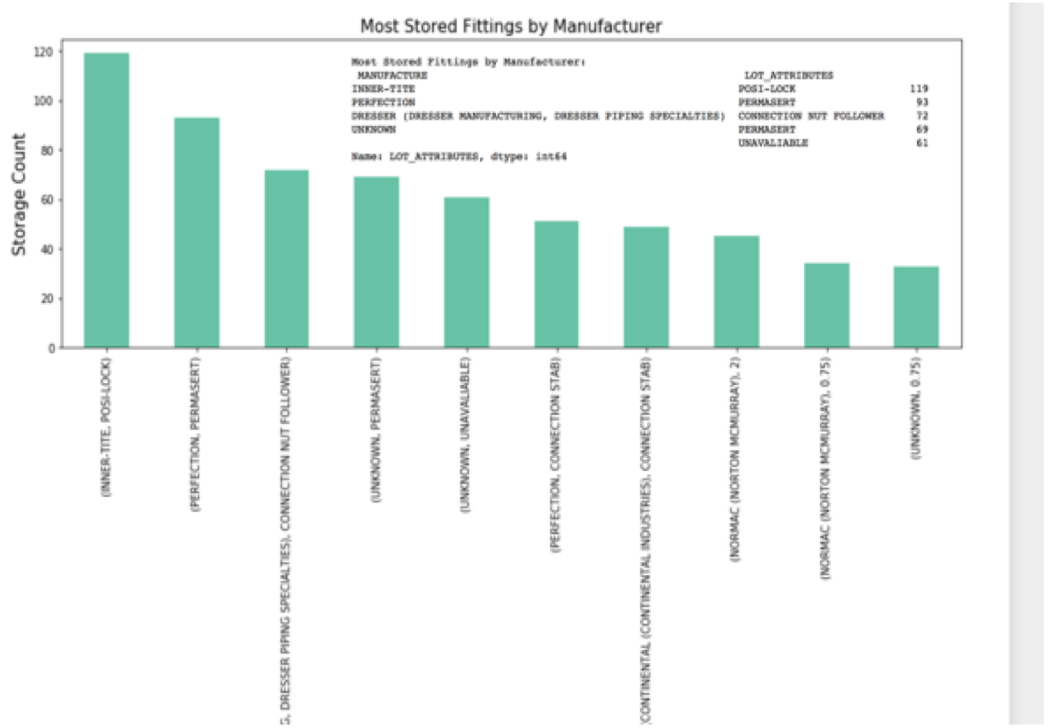


- Highest percentage Unknown

…contd.

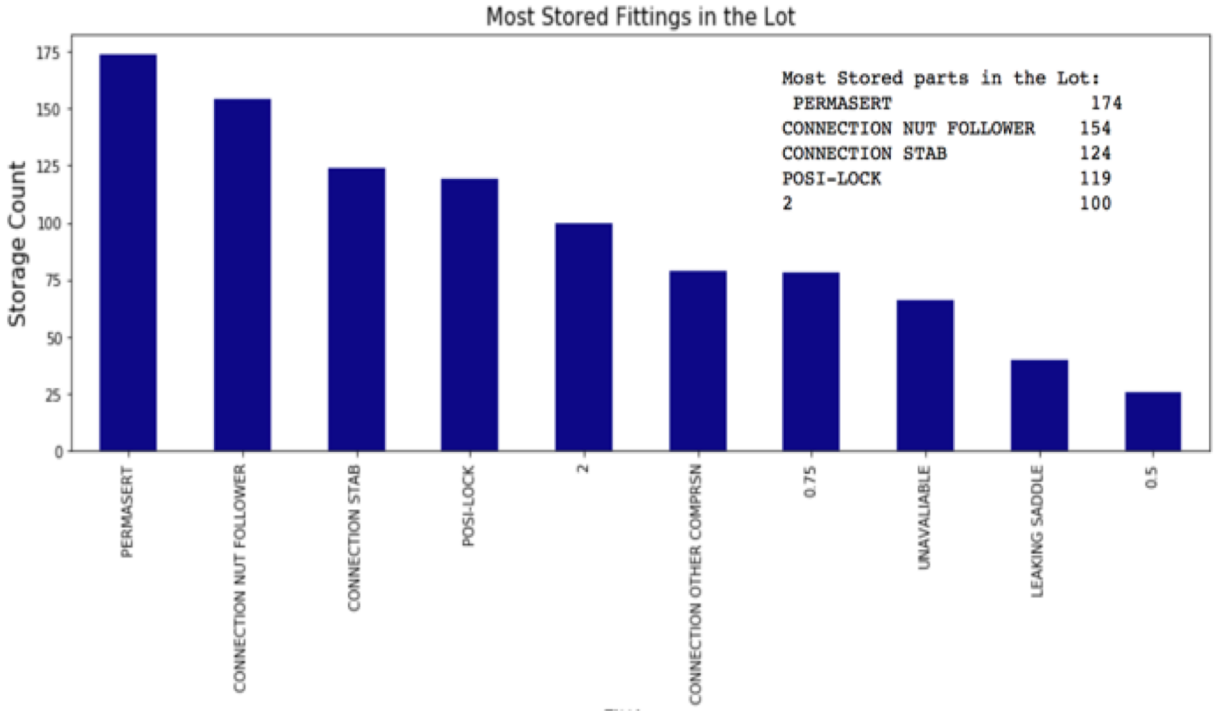
Exploratory Analysis:

Most Stored Fittings by Manufacturer



- Inter-tite carried maximum POSI-LOCK fittings by manufacturer.

Most Stored Fittings by Lot



- Permasert was the most stored fitting in the lot supplied by Dresser and Perfection..

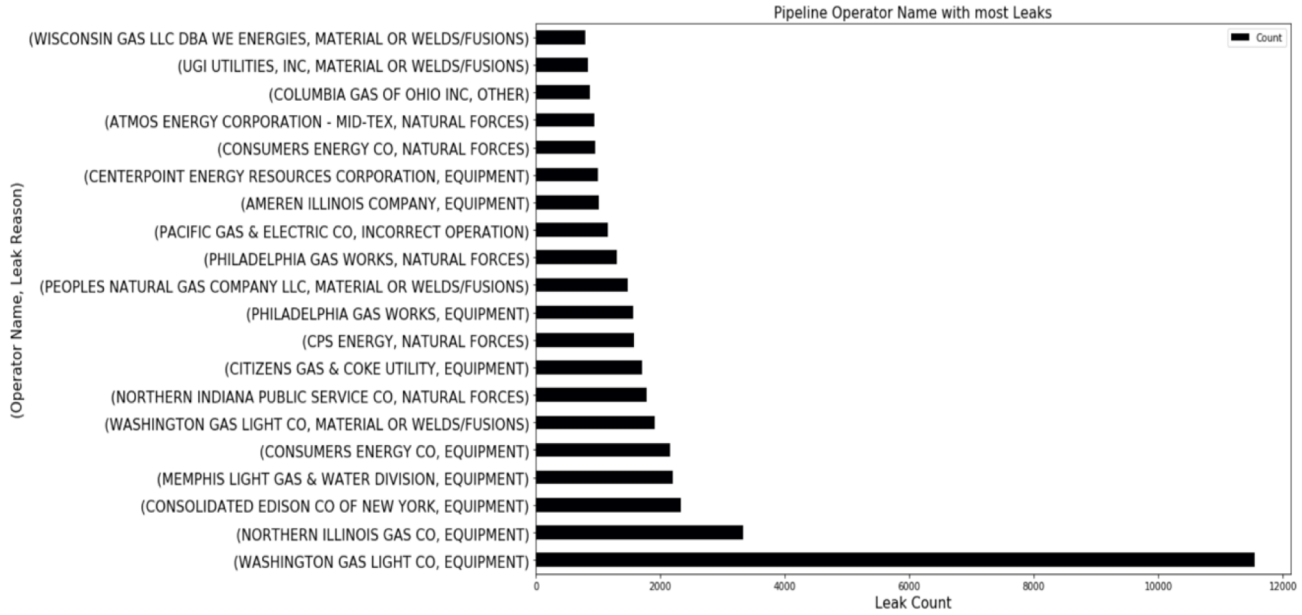
…contd.

Exploratory Analysis:

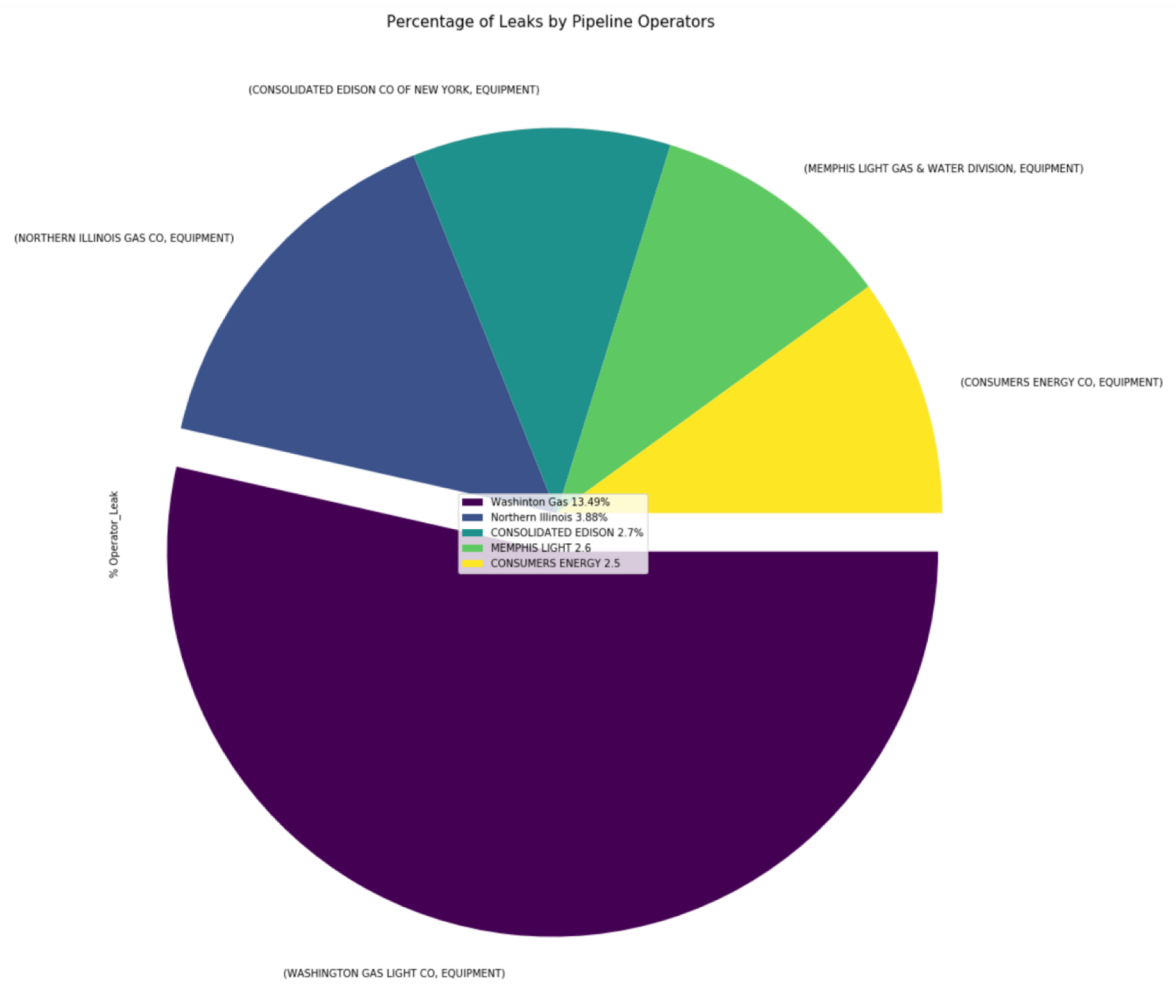
Pipeline Leaks by Operator

Leak Frequency by Pipeline Operator:

OPERATOR_NAME	LEAK_CAUSE_TEXT	Operator_Count	% Operator_Leak
WASHINGTON GAS LIGHT CO	EQUIPMENT	11551	13.492425
NORTHERN ILLINOIS GAS CO	EQUIPMENT	3326	3.885015



Percentage of leaks by Pipeline Operators



Observation:

- Most pipeline operator leaks were due to equipment failure.
- Washington Gas Light and Northern Illinois pipeline operators recorded 13.49% and 3.88% leaks respectively.

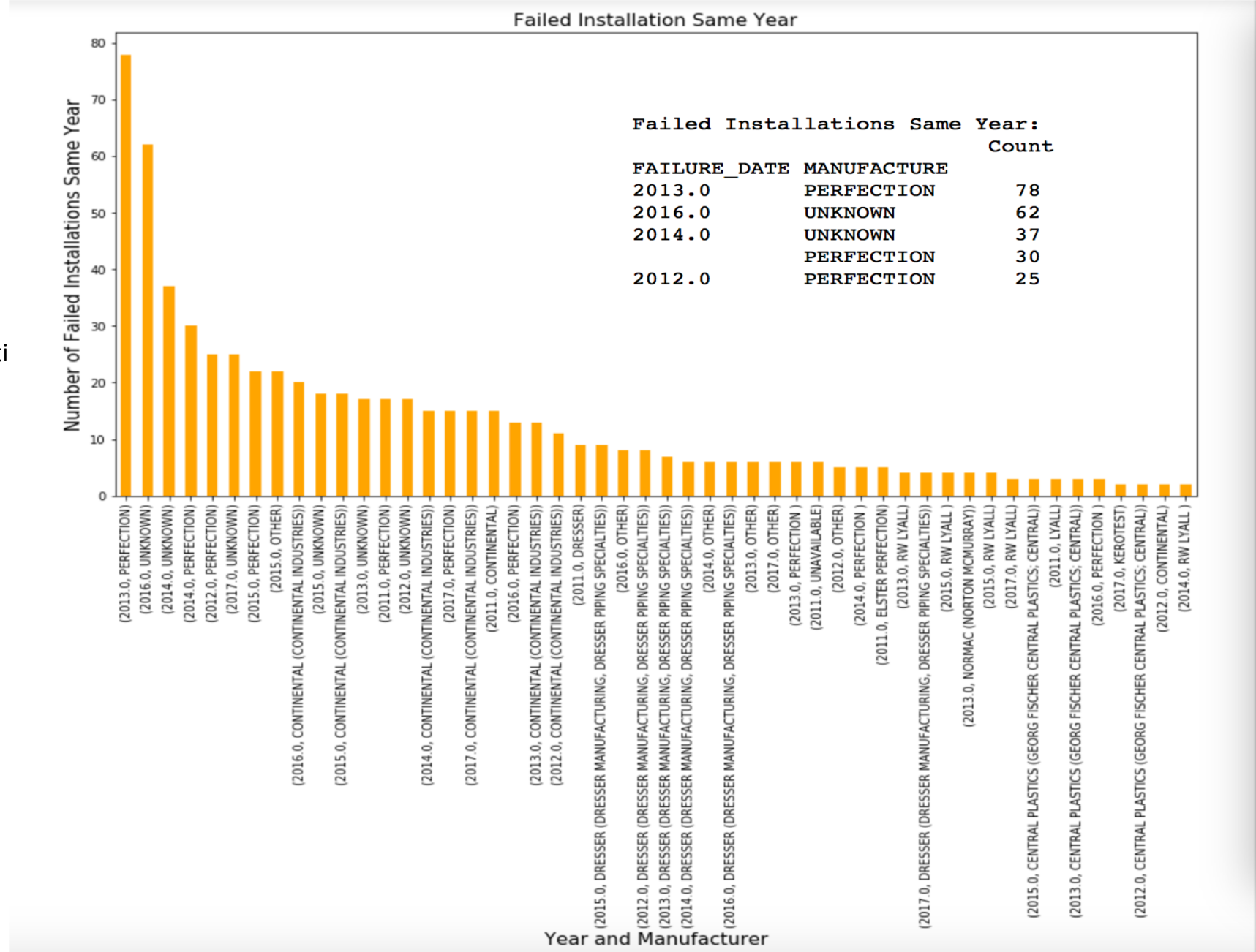
...contd.

Exploratory Analysis:

Same Year Failed Installation

Observation:

- Perfection had maximum number of failed installations in 2013.
- Dresser and Continental and Perfection had multi failed installations between 2011 to 2017.



Machine Learning Process Flow:

Get Data Ready for Modeling

- All Missing Values Removed
- Categorical Values Converted to Numeric
- One-Hot Encoder
  - Binary Encoder

Hyper-Parameter Tuning

- Randomized Search CV
- Grid Search CV

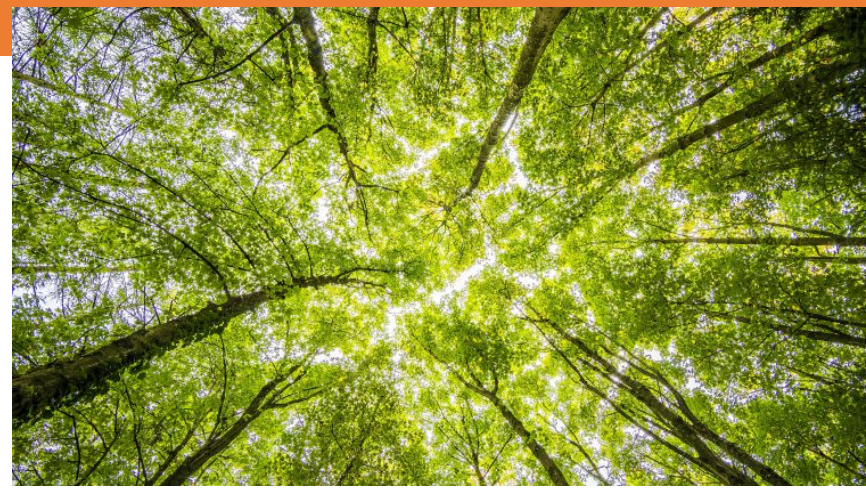
Modeling

- Train Model
- Predict Labels

Evaluate Model

- Classification Report
- Confusion Matrix
- Feature Importance

Random Forest Classifier Hyperparameter Tuning :



- **Randomized Search CV**

Parameters

{'criterion': ['gini', 'entropy'],
'max_depth': [10,30],
'max_features': [10, 20, 100],
'min_samples_split': [2, 5,10],
'min_samples_leaf': [1, 2, 4],
'n_estimators':[100, 300]}

- **Grid Search CV**

Parameters

{'n_estimators': [100],
'min_samples_split': [10, 30],
'min_samples_leaf': [1],
'max_features': [100,140],
'max_depth': [30, 50],
'criterion': ['entropy']}

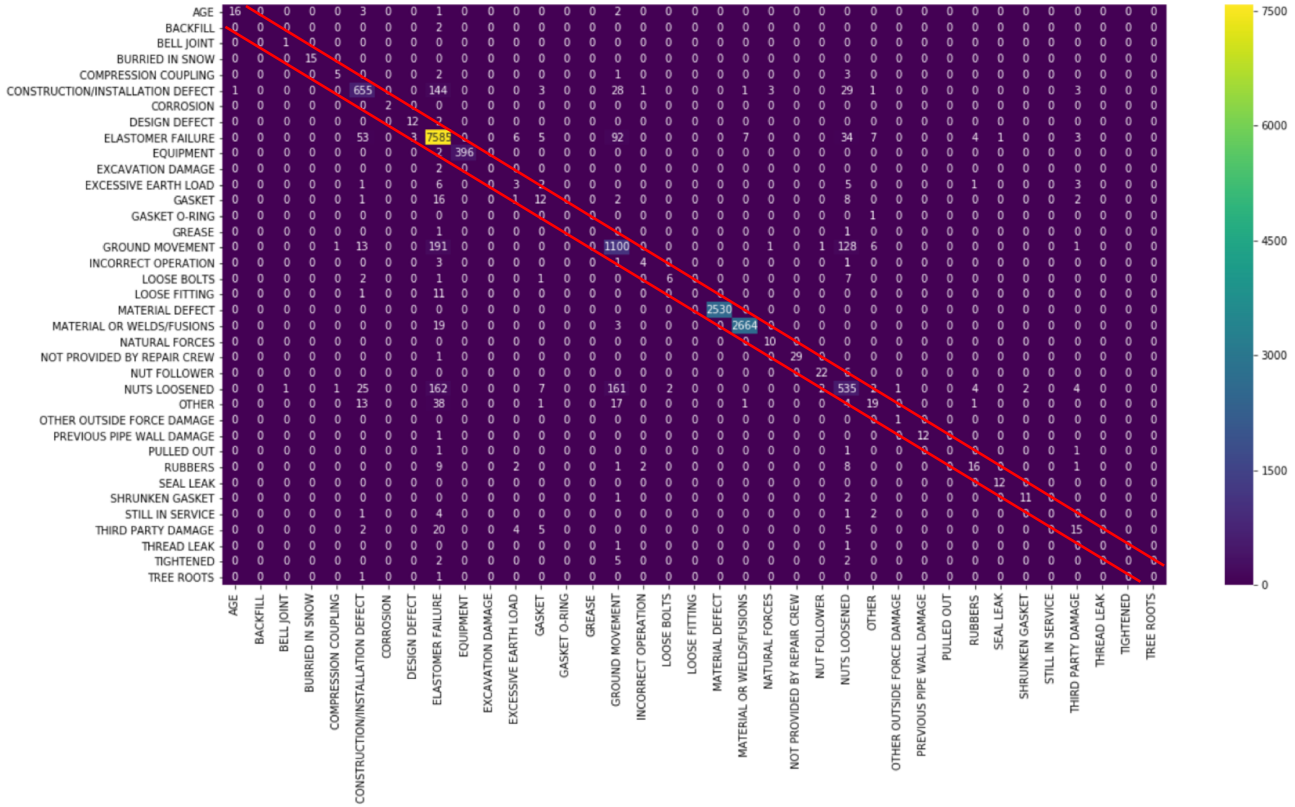
Selected Parameters

{n_estimators= 100,
min_samples_split= 10,
min_samples_leaf= 1,
max_features= 100,
max_depth= 50,
criterion= 'entropy',
random_state= 123}

Classification Report & Confusion Matrix for Model B using Binary Encoder:

Classification Report:

	precision	recall	f1-score	support
AGE	0.94	0.73	0.82	22
BACKFILL	0.00	0.00	0.00	2
BELL JOINT	0.50	1.00	0.67	1
BURRIED IN SNOW	1.00	1.00	1.00	15
COMPRESSION COUPLING	0.71	0.45	0.56	11
CONSTRUCTION/INSTALLATION DEFECT	0.85	0.75	0.80	869
CORROSION	1.00	1.00	1.00	2
DESIGN DEFECT	0.80	0.86	0.83	14
ELASTOMER FAILURE	0.92	0.97	0.95	7800
EQUIPMENT	1.00	0.99	1.00	398
EXCAVATION DAMAGE	0.00	0.00	0.00	2
EXCESSIVE EARTH LOAD	0.19	0.14	0.16	21
GASKET	0.33	0.29	0.31	42
GASKET O-RING	0.00	0.00	0.00	1
GREASE	0.00	0.00	0.00	2
GROUND MOVEMENT	0.78	0.76	0.77	1442
INCORRECT OPERATION	0.57	0.44	0.50	9
LOOSE BOLTS	0.75	0.35	0.48	17
LOOSE FITTING	0.00	0.00	0.00	12
MATERIAL DEFECT	1.00	1.00	1.00	2530
MATERIAL OR WELDS/FUSIONS	1.00	0.99	0.99	2686
NATURAL FORCES	0.71	1.00	0.83	10
NOT PROVIDED BY REPAIR CREW	1.00	0.97	0.98	30
NUT FOLLOWER	0.88	0.79	0.83	28
NUTS LOOSENED	0.69	0.59	0.63	909
OTHER	0.50	0.20	0.29	94
OTHER OUTSIDE FORCE DAMAGE	0.50	1.00	0.67	1
PREVIOUS PIPE WALL DAMAGE	1.00	0.92	0.96	13
PULLED OUT	0.00	0.00	0.00	3
RUBBERS	0.62	0.41	0.49	39
SEAL LEAK	0.92	1.00	0.96	12
SHRUNKEN GASKET	0.85	0.79	0.81	14
STILL IN SERVICE	0.00	0.00	0.00	8
THIRD PARTY DAMAGE	0.45	0.29	0.36	51
THREAD LEAK	0.00	0.00	0.00	2
TIGHTENED	0.00	0.00	0.00	9
TREE ROOTS	0.00	0.00	0.00	2
micro avg	0.92	0.92	0.92	17123
macro avg	0.55	0.53	0.53	17123
weighted avg	0.91	0.92	0.91	17123



Trained model classified 38 reasons that caused Mechanical Fitting to Leak in the gas pipeline.

Accuracy Score: 91.6%

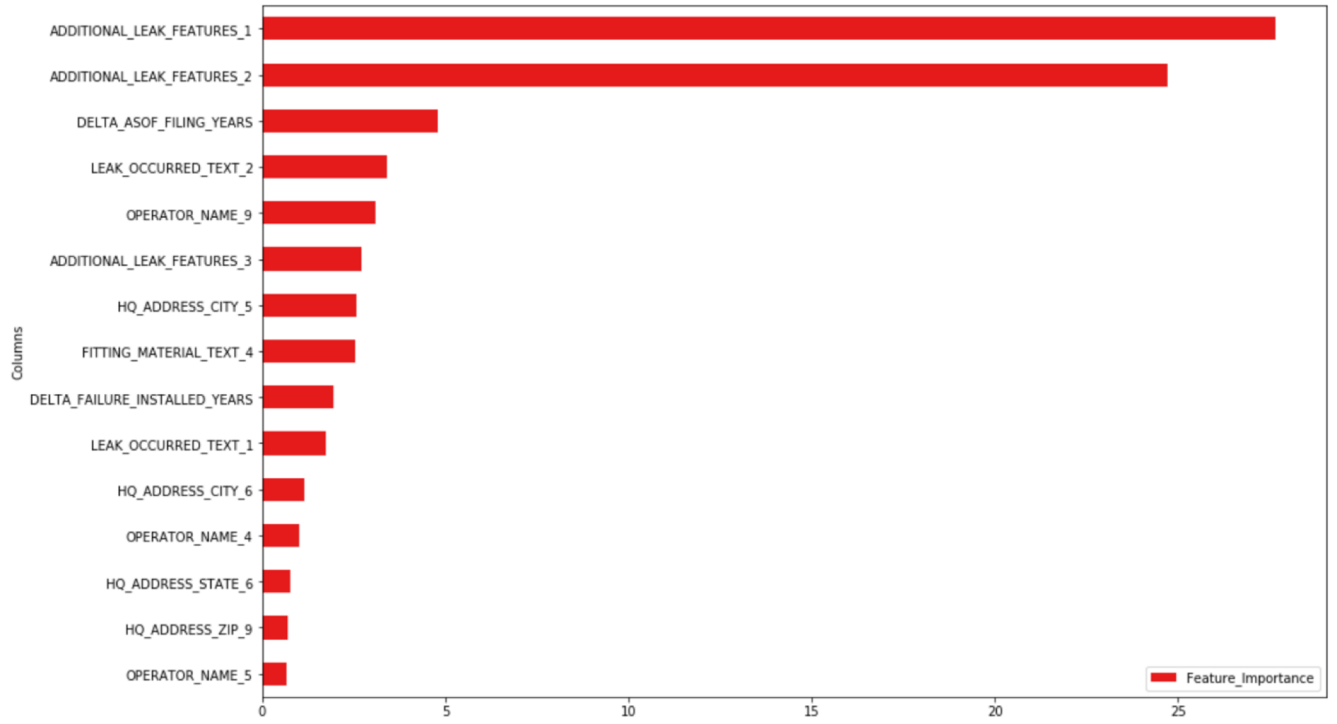
Feature Importance:

Top 15 Features used in Classifying Leak Cause

```
Top 15 Features for Classifying Leaks in Pipeline:
 Feature_Importance    79.525348
dtype: float64
```

Columns	Feature_Importance
ADDITIONAL_LEAK_FEATURES_1	27.665679
ADDITIONAL_LEAK_FEATURES_2	24.703340
DELTA_ASOF_FILING_YEARS	4.783924
LEAK_OCCURRED_TEXT_2	3.396982
OPERATOR_NAME_9	3.096549
ADDITIONAL_LEAK_FEATURES_3	2.719892
HQ_ADDRESS_CITY_5	2.588522
FITTING_MATERIAL_TEXT_4	2.554088
DELTA_FAILURE_INSTALLED_YEARS	1.934990
LEAK_OCCURRED_TEXT_1	1.751176
HQ_ADDRESS_CITY_6	1.153430
OPERATOR_NAME_4	1.029005
HQ_ADDRESS_STATE_6	0.767364
HQ_ADDRESS_ZIP_9	0.716673
OPERATOR_NAME_5	0.663734

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8403617240>
```



Conclusion:

- Random Forest using Binary Encoded Data (Model B) classified top 15 causes for leak with 91.6% accuracy.
- Random Forest using one-hot encoded data (Model A) classified top 15 causes for leak with 90% accuracy.
- Model B run time was faster than Model A due to significant difference in column sizes (136 Vs 3,912)
- Model B used top 15 features to predict 80% of the leak classification while Model A classified only 48.8%.
- Random Forest using Binary Encoder (Model B) was preferred.
- Important Features which classified leak cause correctly were:
    - Excavation damages
    - Natural forces
    - Welding defects
    - Leak Through the Body or Seal
    - Time elapsed between 'Installation' and 'Failure Date'.
    - Time elapsed between 'As of' and 'Filing Date'
    - Pipeline Operator

Future Improvements:

Existing Model:

- Avoid mixed classification in the same class
  - Elastomer Failure 7,642 examples
  - Construction/installation defects 144
  - Ground Movement defects 191
  - Nuts Loosened defects 162

- Gather Data pertaining to Equipment Failure.

Apply Other Advanced Models:

- Use deep learning to classify leak cause
- Forecast mechanical fitting failure date based on years of operation.

Build Dashboard:

- Provide real time information on key performance indicators .
  - Expected failure date
  - Operator and Manufacturer information
  - Reason for leak cause